The `Dataset` project: handling survey data in `R`

Emmanuel Rousseaux, Gilbert Ritschard
Institute for Demographic and Life Course Studies, University of Geneva
NCCR LIVES: Overcoming vulnerability: life course perspectives

Motivation

- Currently no real standard for sharing survey data in `R`
- Need for consistent weight handling
- Manipulating longitudinal data still difficult
- A lot of state-of-the-art methods are provided in `R` only

Goals

- Storing efficiently survey data
- Specific design for longitudinal data
- Assist the user on the pre-processing steps for a specific analysis
- Exporting directly data and user manual for sharing

Proposal

- 2 packages in `R`
  - `Dataset`: for cross-sectional survey data
  - `stDataset`: for spatio-temporal survey data
- Full S4
- `https://r-forge.r-project.org/projects/dataset/`

Overview

General specifications

- Sophisticated managment of missing values
- Automatic consistency tests
- User-oriented functions
- Automatic summaries

Entering and storing data

Specific methods for entering

- cross-sectional data
- longitudinal data
- network data

Pre-processing data for a specific study

- Efficient recoding operations
- Use information from user manual ⇒ `contains(x, 'health')`

Representativity is central. Efforts have to be made for helping the user

- Real structure for handling weights in the database
- Representativity checks on each variable
- Generate new weights to correctly balance a subdataset

Exporting data for sharing

```
export(mydataset, file = 'myfile.RData')
```

Provide a full description file in PDF ⇒ ready for sharing

Facilitate rendering of spatial Data

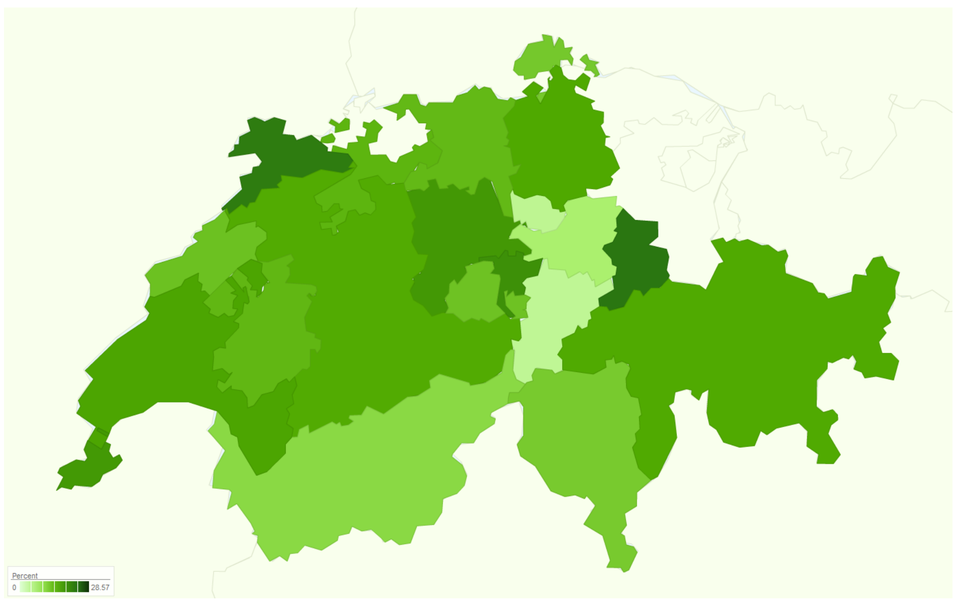

Figure 1: Poor/Good SRH ratio. PSM 2011, wave 2010 (no weighted)

Interfaces for common analysis tools

- Bivariate analysis (Cramer's V, Kendall's tau a, Theil's u, Somer's D, ...)
- Logistic regression
- ...

Exports in easily readable PDF files.

Design

`Variable` object

The `Variable` object is represented by

`codes`: vector of codes
`missings`: vector coding/labelling missing values
`values`: vector coding/labelling valid cases
`description`: variable label

Specific `Variable` types

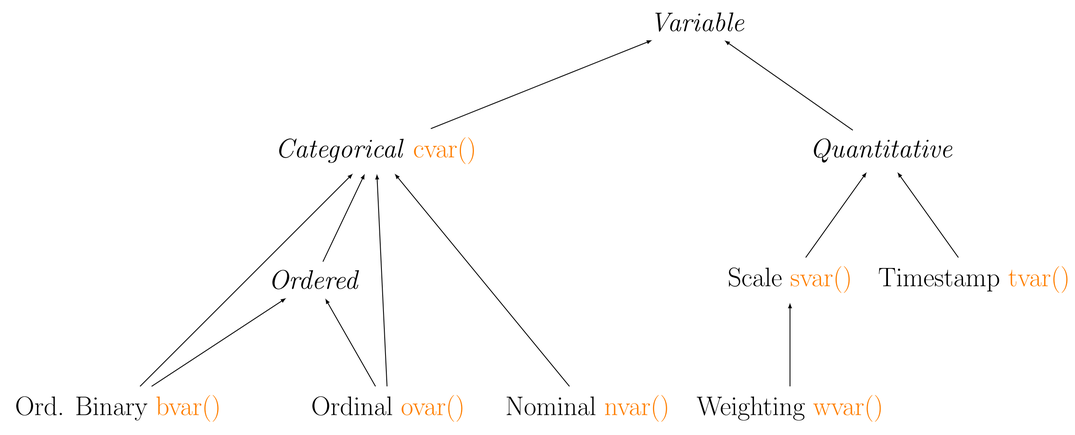

Figure 2: Class diagramme of objects inheriting of the Variable class. Virtual classes are in italics.

`Dataset` object

The `Dataset` object is represented by

`variables`: list of variables
`name`: name of the dataset
`description`: a long label
`row.names`: names for rows
`spatial`: a variable used as spatial reference
`weighting`: a variable used for weighting
`control`: some control variables
`infos`: a list for storing other information the user want to share

Creating a `Dataset` object

From an existing `data.frame`
From a SPSS file
By hand (as a list of `Variable` objects)

Toward "Life course" objects

- `stDataset`: design based on `Dataset`
- Longitudinal summary (in PDF)
- Storing, extracting and manipulating trajectories directly
- Construction of a "life course" object ready for analysis

Forthcoming presentations

- (SRSSS, Unige) Manipulating panel data with `stDataset`
- (SRSSS, Unige) Handling weights with `Dataset` and `stDataset`, illustration with the Swiss Household Panel

About me

PhD Position

Teaching and Research Assistant at the Departement of Economics, SES, Unige. PhD directed by Gilbert Ritschard (iDemo, Unige) and Michel Léonard (ISS, Unige). Participating to the NCCR LIVES as member of IP14: "Measuring life sequences and the disorder of lives" leaded by Gilbert Ritschard.

Overview of the thesis project

- Providing a software framework for handling survey data (in `R`)
- Providing a software framework for handling life courses as a whole
- Providing new mining tools for rare events
  - Decision trees for the discovery of vulnerable profiles
  - Multi-channel association rules mining
- Apply these tools for getting new insights on poor health situations

Acknowledgement

This publication results from research work is executed within the framework of the Swiss National Centre of Competence in Research LIVES, which is financed by the Swiss National Science Foundation. The authors are grateful to the Swiss National Science Foundation for its financial support.